DOCUMENT RESUME

ED 221 044 FL 013 159

AUTHOR Reid, Joy M.; O'Brien, Maryann
TITLE The Application of Holistic Grading in an ESL Writing Program.
PUB DATE Mar 81
NOTE 22p.; Paper presented at the Annual Convention of Teachers of English to Speakers of Other Languages (15th, Detroit, MI, March 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *English (Second Language); *Evaluation Criteria; *Holistic Evaluation; Language Tests; *Scoring; Test Reliability; Test Validity; *Writing (Composition); *Writing Evaluation

ABSTRACT

Holistic evaluation, defined as the grading of an entire piece of writing without tallying or correcting specific errors, is discussed as preferable to discrete point testing or objective measures in assessing writing proficiency of English as a second language (ESL) students. The assessment process, in which the papers are evaluated as excellent, good, average, fair, or poor in each of three areas (English structure, content, and organization) is described. Some preliminary research results in using holistic evaluation for ESL writing are reported and the method is found to be a reliable and valid measure. The scoring sheet and evaluation guide are appended. (EKN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED221044

The Application of Holistic Grading in an ESL Writing Program

Joy M. Reid and Maryann O'Brien

Despite the reservations of some proponents of discrete point testing, an essay sample is often the clearest example of an ESL student's proficiency in writing. The problems of objectivity and validity exist; however, if the raters are well trained and the criteria important in making the evaluation are clear, the reliability increases significantly. This paper presents a set of criteria and a scoring sheet adapted from both ESL and native speaker evaluative scales for writing samples. The process for evaluation and ranking is discussed, initial research concerning the resulting data is presented, and the applications of the procedure within an ESL writing program are described.

FL 013159

The problems of testing and assessing a second language student's writing proficiency are varied and difficult to solve. Discrete point tests and other objective techniques cannot accurately measure this aspect of the student's competence (Carroll 1973, Oller 1976). Objective writing tests "do not allow one to see how a second language learner organizes his thoughts on paper or applies his known vocabulary, nor do they indicate how well a learner uses in extended, unified prose the formal

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Joy Reid
Maryann O'Brien

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

grammatical rules he has been taught . . . ." (Mullen 1977). There is therefore a need for an integrative measure of the student's writing; the test involved is usually a free writing sample: the student is asked to write a paragraph/essay on a specified topic or picture. The results of such a test can be used for program placement, class evaluation, pre- and post testing, and final class grades.

Assessing the writing sample presents significant problems in reliability and validity because of the subjective quality of the evaluation. These problems are compounded in an ESL writing program associated with an MA TEFL program in which several teachers with varying backgrounds in composition analysis and teaching experience are involved in evaluating student writing. This paper explains one solution to the problem of assessment: the development of a well-planned scoring guide for holistic evaluation of writing samples and the training of teachers in its use.

Holistic Evaluation

In contrast to discrete point testing and objective measures of writing proficiency (Perkins 1980), holistic evaluation is described as a procedure for scoring or ranking a piece of writing wholly, without tallying or correcting specific errors. Teacher training is the significant difference in both individual outcome and overall reliability between simple intuitive/subjective essay grading and holistic evaluation.

> The placing, scoring or grading occurs quickly, impressionistically, after the rater has practiced the procedure with other raters . . . . Holistic evaluation is usually guided by a holistic scoring guide which describes each feature and identifies high, middle, and low quality levels for each

feature.

> When there is commitment and time to do the work required to achieve reliability of judgment, holistic evaluation of writing remains the most valid and direct means of rank-ordering students by writing ability (Cooper 1977).

Development of the Evaluation Guide

The first step in the holistic grading process, then, is the development of a grading guide which is clear and easy to use, parallel in structure, and inclusive of the skills taught in the ESL writing program. This evaluation guide should organize and sensitize the raters' perceptions of writing and provide criteria on which to base their grading. At Colorado State University's Intensive English Program (IEP), the initial holistic scoring guide was developed from three sources (Mullen 1977, Powell 1980, Reid 1980). The first two were ESL scoring sheets for composition grading, the third a grading guide used at CSU in the Freshman Placement Testing Program. The latter originated from ETS Advanced Placement grading guides and was modified for use at CSU. The criteria in each of the guides were similar and consisted of evaluation in three major areas: organization (the process of thought and support), content (the quantity and quality of thought and diction), and English structure (the use of grammar and mechanics). The specific features and weight given in these areas differed, depending on the overall goals of the guide developers. For the IEP, the analytic scale was modified during several sample evaluation sessions by the writing teachers in the program. Specific modifications emerged as a result of a refining process which was in itself a teaching/learning process for the teachers. (See Appendix A for the

grading guide). At the same time, the teachers developed a 9-point evaluation scale with five main levels: excellent, good. average, fair, and poor. (See Appendix B for the evaluation scale). The flexibility of being able to mark on the lines as well as in the spaces proved an advantage for the teachers in the scoring process.

The Writing Sample

A writing sample was administered in all writing classes on the same day. Subjects included 75 students in the IEP who were enrolled in 7 levels of proficiency from beginning to advanced writers. Students were given a class period (approximately 45 minutes) to complete the sample. Students were given only a single topic to write about; sample topics administered during the past year include the following:

1. Write about something that has happened to you in the United States which has made you happy OR angry OR sad.
2. Discuss the difference(s) between going shopping for food in your country and going shopping for food in the United States.
3. Discuss the difference(s) between you and your grandfather (or grandmother).

The samples were collected and names deleted; 9-12 samples were selected by the table leader(s) for the standardizing process and xeroxed to provide enough copies for all the raters. The table leader(s) then read and evaluated the standardizing samples; notes concerning the samples were assembled for the process of calibrating the raters.

Standardizing

Below is a typical group of raters in the IEP:

experienced writing teachers	2
experienced teachers in other ESL areas	2
graduate teaching assistants with some ESL experience in teaching writing	2
graduate teaching assistants with no experience in teaching ESL writing	2
table leaders	2

The actual standardization process consisted of the following steps. First, the principles of holistic grading were explained and discussed by the table leaders as well as how the process of the holistic evaluation session would occur. Then the scoring guide was distributed, explained, and thoroughly discussed. Following, the scoring sheet was distributed and discussed. Finally, the raters broke into small groups (4-5 per group) to standardize; each group was headed by a table leader. The raters were each given a group of three writing samples and asked to read the samples, evaluate them in terms of the scoring guide, and mark the scoring sheet accordingly. When each group had finished, the table leader asked for individual scores. Discussion of these scores followed until unanimity or near unanimity (within 1 point on the scoring sheet) resulted. This process was repeated two or three times, depending on the time allotted, with each set of standardizing papers presenting increasingly difficult problems that further refine the perceptions of the raters:

Set 1 a clear low level (1-2)paper
a clear intermediate (5) paper
a clear advanced (8-9) paper

Set 2 a 3-4 paper

an 8 paper

a 4-5 paper

Set 3 a 2-3 paper

a 5-6 paper

a 6-7 paper

(See Appendix C for a typical Set 1).

Evaluation

Raters assembled at a long table (or tables, each with a table leader). Each rater was given a stack of randomly selected writing samples; using the evaluation guide, the rater read and evaluated a sample, scored it on the scoring sheet, put the scoring sheet on the back of the sample and attached a clean scoring sheet to the front for a second evaluation. As the second readings progressed, the table leaders checked the scoring sheets and determined which sampled had been evaluated unanimously (i.e., within one point of the scoring average at the bottom of the scoring sheet) and which samples needed a third evaluator. Following are typical figures for an evaluating session:

standardization: 1 hour (25 minutes for discussion; 35 minutes for sample evaluation)

evaluation: 1 hour

75 papers double-read in 50 minutes

15% of those papers need and receive a third reading in 10 minutes

Sample Data

Below is a sample page of raw data collected during a single term:

Table 1

Raw Data Sample

Subject	TOEFL	Class	Class Ranking	Evaluation	Evaluation Ranking	Country
A	513	56	1/13	6-6	7/12	Venezuela
B	350	45	4/13	3-4-5	6/12	Honduras
C		36	11/14	2-3-4	10/12	Saudi Arabia
D		26	2/10	1-2	3/7	Columbia
E	460	66	7/12	8-9	4/12	South Vietnam
F	486	56	7/13	6-7	5/12	Yemen
G		45	5/13	3-3	9/12	Venezuela
H	507	66	9/12	7-7	6/12	Taiwan
I		56	6/13	7-7	2/12	Thailand
J		36	4/14	4-5	2/12	Japan
K		46	11/14	4-5	9/14	Saudi Arabia

Research

Obviously such raw data can provide opportunities for a variety of research. Following are two initial studies within the IEP/MA TEFL program at CSU. The first was designed to test the following hypothesis: after explanation of the grading guide and practice in its use, both inter-rater and intra-rater reliability would increase.

Subjects: 10 MA TEFL candidates, all native speakers of English who had had no previous experience in teaching ESL writing.

Procedure: a writing sample was administered throughout the writing classes in the IEP as described earlier in this article, and two table leaders selected and evaluated standardization samples.

The subjects assembled and evaluated the same eight samples individually without prior knowledge of holistic grading or the grading guide. Then the table leaders provided the same standardizing process described earlier. Finally, the subjects were given one hour to read and evaluate fifteen samples. In order to test intra-rater reliability as well as inter-rater reliability, the fifteen samples consisted of the following:

3 samples from the first grading

5 new samples, each duplicated once and placed at varying distances within the sample packets

2 new samples appearing only once

Research questions addressed:

1. To see whether there was any significant difference between

the rating of the three repeated samples graded before and after explanation of the grading guide, three paired t-tests were done.

Results: critical value of the t at the .05 level is 1.812

t = 1.892

t = .93

t = .145

The difference in the way the raters graded the same samples before and after explanation of the grading guide was significant in only one case.

2. To test intra-rater reliability, the mean of each grader on the first appearance of the five repeated samples was compared to their mean on the second appearance of these samples.

Results:		mean on the first appearance	mean on the second appearance
Subject	1	5.22	4.80
	2	4.48	4.20
	3	4.54	3.94
	4	4.46	4.74
	5	5.12	5.40
	6	4.14	4.36
	7	4.34	4.40
	8	4.46	5.12
	9	3.40	3.54
	10	4.00	3.80

Each rater's mean differed by less than one point between the two repeated samples; therefore, raters were fairly consistent

in their individual evaluation of a paper.

3. To measure whether or not there would be an increased level of consistency among the raters after they had used the grading guide, the variances before and after were compared.

   Results: before treatment $s^2 = .656$ $s = .809$

   after treatment $s^2 = .260$ $s = .509$

   The variability was therefore smaller after treatment; the raters were more in accord with each other.

The second study was designed to test the following hypothesis: experienced writing teachers in the IEP would be more consistent than beginners.

Subjects: 7 experienced writing teachers in the IEP; experience ranged from 1 term of beginning writing to several years of teaching writing in the program.

Procedure: using the same materials as in the preceding study, the subjects were standardized and then given an hour to evaluate the fifteen packet samples.

Research questions addressed:

1. To test intra-rater reliability, the mean of each rater on the first appearance of the five repeated samples was compared to their mean on the second appearance of these samples.

   Results:

		mean on the first appearance	mean on the second appearance
Subject	1	5.0	5.3
	2	3.9	4.3
	3	5.3	5.3

4	4.5	3.8
5	4.9	4.9
6	4.5	4.5
7	5.2	5.2

As with the first group, each rater's mean differed by less than one point between the two repeated samples; therefore, raters were fairly consistent in their individual evaluation of a paper. However, four of the seven raters achieved exactly the same mean, indicating a high standard of intra-rater consistency.

2. To measure whether or not there would be an increased level of consistency among the experienced teacher-raters when compared with the first group, the variances of the two groups were compared.

Results: first group before treatment $s^2 = .656$ $s = .809$

first group after treatment $s^2 = .260$ $s = .509$

second group $s^2 = .187$ $s = .432$

While not significantly different, these figures indicate a definite trend towards consistency.

Based on the preliminary data above, and allowing for the small N size, it would seem that the more experienced the raters are in teaching ESL writing and/or the more practice the raters have with holistic evaluation, the higher both the inter-rater and the intra-rater reliability will be.

Conclusions

Within the narrowest and least scientific scope, the integrative testing and holistic grading sessions in the CSU IEP writing program were beneficial in the following ways:

1. the evaluation scores of the writing samples generally corresponded to the placement of students and to the writing skills taught at different levels in the program.
2. the available TOEFL scores showed a correlation with student placement and with the writing samples (with the possible exception of #4 below).
3. the data relating teacher ranking to writing sample ranking provided the opportunity for several observations:
    A. it gave all the teachers a new perspective on some students.
    B. it showed that there is often a significant difference in student performance between in-class writing and home assignments.
    C. it indicated that students may need more practice writing in class under the pressure of time.
4. at the upper levels in the writing program, the initial writing samples were not as valid a test of skills, probably because in order to include low level writers the assignment was often narrative or descriptive while the concentration in the intermediate and advanced levels in the IEP writing program is on exposition and analysis. Therefore, in subsequent testing, writing

tasks need to be selected which allow for more valid discrimination of skills at the upper levels.

5. the evaluation process gave the writing teachers an opportunity to discuss the features and specific skills taught in the IEP writing program, to agree upon approximate levels of competency, to learn about the capabilities and limitations of other students in the program as well as their own, and to share very specific opinions and classroom experiences with other writing teachers. Such an opportunity is valuable for all teachers and is essential for new and less experienced teachers in that skill area.
6. finally, the evaluation process allowed the participants to see that grading can be done holistically and still be both reliable (approximately 15% of the papers had to be read a third time for uninimity, and none had to be read a fourth time) and valid (i.e., the writing could reflect what the students were learning in the IEP writing program).

In its applications to other ESL writing programs, integrative testing and holistic evaluation can be useful in several ways:

1. to indicate whether or not existing placement procedures in writing are valid.
2. to serve as a writing placement tool.
3. to indicate whether or not students are being evaluated consistently throughout the writing program.
4. to indicate whether or not the separate levels within a writing program are providing the writing skills that the students must acquire.

5. in the area of teacher training, it can

   A. bring experienced teachers together to discuss evaluation procedures.

   B. teach inexperienced teachers about the evaluation of student writing.

Certainly there is still an enormous amount of research to be done in the areas of integrative testing and holistic evaluation in ESL writing programs and in ESL programs in other skill areas as well. The following represent some areas of research available from the raw data shown above:

1. the examination of objective measures such as error-free T-units, the Larsen-Freeman index of development, or the Flahive-Snow complexity index and their possible correlation with holistic scoring.
2. the comparison of holistic scores with standardized measures like the TOEFL test or the Michigan test.
3. the comparison of teacher-ranking of students within a class and holistic evaluation and ranking of student writing samples within a class.
4. continued examination on the increased reliability of more experienced raters in holistic evaluation.

Researchers working in the area of holistic evaluation would be well advised to consider the studies already completed with native-speaker papers. Following such research, Cooper (1977) concluded that "holistic evaluation can be as reliable as multiple-choice testing and . . . it is always more valid." The use of holistic evaluation guides should therefore be used in the placement, prediction of success in English, and growth measurement of specific skills in ESL writing programs.

References

Cooper, Charles R. 1977. Holistic evaluation of writing. Evaluating writing. New York: NCTE: 3-32.

Carroll, John B. 1973. Foreign language testing: persistent problems. Testing in second language teaching: new dimensions. Maureen Concannon O'Brien (Ed.). Dublin: ATESOL-Ireland: 6-17.

Mullen, K. A. 1977. Using rater judgments in the evaluation of proficiency for non-native speakers of English. On TESOL '77: 309-320. H. D. Brown, C. A. Yorio and R. H. Crymes (Eds.).

Oller, John. 1976. Language testing today: an interview with John Oller. English Teaching Forum (July 1976): 22-27.

Perkins, Kyle. 1980. Using objective methods of attained writing proficiency to discriminate among holistic evaluations. TESOL Quarterly 14, 1: 61-69.

Powell, Bill. 1980. ESL and composition: a report. TESOL newsletter, XIV, 3: 15-16.

Reid, Stephen. 1980. Grading guide developed for use in evaluating Freshman Placement Examinations at Colorado State University.

APPENDIX A

EVALUATION SHEET

ENGLISH STRUCTURE

Excellent	Few noticeable errors in grammar, word order, punctuation and/or spelling; frequent use of complex sentence structures.
Good	Occasional errors in grammar, word order, punctuation and/or spelling; some use of complex sentence structures.
Average	Moderate errors in grammar, word order, punctuation and/or spelling; general use of simple sentences; occasionally an obscured meaning.
Fair	Frequent errors in grammar, word order, punctuation and/or spelling make comprehension difficult; use of short basic sentence structures.
Poor	General use of phrases and/or fragments; very basic vocabulary and/or rampant misspelling.

CONTENT

Excellent	Writing is an easy task; quantity is no problem; diction is broad and appropriate; interesting, substantial, complex ideas that respond directly and completely to the assignment; topic developed fully and coherently.
Good	Writing flows without much hesitation; reasonable quantity; some diction errors of complex words; interesting ideas that are relevant to the assignment; ideas develop the topic with specific detail and coherence.

Average	Evidence of having stopped writing at times; somewhat limited vocabulary; adequate development of ideas relevant to the assignment with some specific detail; flawed coherence.
Fair	Limited quantity; limited development; little specific detail; simple ideas not necessarily related to the assignment; simple vocabulary with frequent diction errors.
Poor	Little writing; very simple ideas, often not directly related to the assignment; very basic vocabulary with frequent errors.

ORGANIZATION

Excellent	Clear introduction directed in an interesting way towards a specific audience; paragraph(s) developed linearly; topic sentence, logical support with substantial specific data; successful use of coherence devices within sentences and between sentences; clear conclusion.
Good	Obvious beginning/middle/end; division of central idea(s) into smaller parts; awareness of audience; paragraph(s) flawed in development and/or logic coherence within sentences but some choppiness; conclusion often a simple restatement.
Average	Intent to develop central idea; some sense of beginning/middle/end; occasional wandering from topic; some development of ideas; general statements predominate; limited use of coherence devices, especially between sentences; conclusion simply stated or missing.

Fair	Limited organization beyond the sentence level; thoughts written down as they come to mind; no introduction or conclusion; coherence limited to internal sentence connectors.
Poor	No apparent organization; no introduction or conclusion; no focus; little or no development; no coherence devices.

APPENDIX B

SCORING SHEET

SUBJECT _____

Evaluator _____ Date _____

	Poor	Fair	Average	Good	Excellent
English Structure					
Content					
Organization					

Overall Writing Proficiency									
	1	2	3	4	5	6	7	8	9

Average Score _____

Appendix C

Set I: Student Samples for Standardization

Subject A

The shopping in my country vary match, thier in more food farch. but not thich differenc. jast clus the stor same. the paopl went at the Mosc.

Subject B

I came from China. There are many differences between going shopping for food in China and going shopping for food in the U.S. I almost went shopping for food in China every day. In the store, there is a counter in front of the food shelf. Usually, the people have to pay the food first, and then pick up them.

In the U.S., the food shelves are showed out of the counter. The people can pick up the food that you want, then pay them.

I prefer going shopping in America.

Subject C

There are differences between going shopping for food in the U.S. and in my country, Columbia. If you are in Bogota, the capital of my country, there are not many differences between going shopping there and in the U.S. because there are supermarkets in Bogota just like in the U.S. But if you go shopping in the little villages of Columbia there you can find many differences. In the U.S. you go shopping for food to a big and organized supermarket where almost everything is canned like juices and fruits. In a little town in Columbia there is not such a supermarket; there is a

special place called plaza de mercadu which doesn't have roof or walls; it doesn't have any kinds of shelters. There the fruits, vegetables, meat, etc. are on tables where the people can touch and see them to know if they are fresh or in bad conditions. That's a difference, because in the U.S. you cannot touch and see the thing that you are going to buy.